Exercise: Big data characteristics

Instructions for completing the exercise

You will find the detailed instructions on how to complete this exercise in the Jupyter notebook file big_data.ipynb . To start with, you need to download the synthetic data set dataset.json and the file containing the users' information users4b.json and upload them to the directory that contains the notebook file, so that the code block in the notebook can read the data properly.

All the points in this exercise will come from answering the questions on A+, but for some of them you need to extract information from the data set by code. Your implementation will not be assessed; the idea is for you to reach the right conclusion no matter how you implement the analysis.

Exercise description

You have started working as a data science intern at a new and exciting Social Media Company (SMC). On SMC's platform (SMP), users can connect to other users via uni-directional follow links, they can make short posts of at most 280 characters, mention other users in their posts, and the posts made by a user are shown in the feeds of all the users following them. Additionally, users can engage with other user's posts by reposting or replying to them or by liking or commenting on them.

In your first project at SMC, you are collaborating with World Climate Organisation to use SMC data to gain more insights about the ongoing climate debate. For this purpose, you have gotten access to a data set of SMC posts discussing climate.

Before diving in to the data, your supervisor reminds you to do preliminary analysis to investigate whether the dataset demonstrates big data characteristics. Therefore you set out to examine and reason about each characteristic by answering the following questions.

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Big

Question 1 10 / 10

To assess how big this "big data" dataset is, compute the average number of original posts made on SMP per day.

- Less than 10,000
- **Between 10,000 and 50,000**
- Between 50,000 and 100,000
- More than 100,000

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Always-on

Question 1 10 / 10

Next, calculate how long is the maximum gap between any two consecutively collected posts in the SMC data set to understand it's always-on nature.

- Less than 1 minute
- **Between 1 and 2 minutes**
- Between 2 and 5 minutes
- More than 5 minutes

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Non-reactive

Question 1 10 / 10

When a person is using SMP, they are not typically aware that their behavior may be observed by researchers and data scientists. So, do you think their posts precisely reflect their behavior or attitudes towards climate debate?

- Yes, they always disclose their unfiltered opinion on SMP since they are not aware of being observed by the researchers.
- **No, even if they have opinions on climate change, they may choose to not voice them on SMP due to social desirability bias.**

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Incomplete

Question 1 10 / 10

Now that you want to use the SMC data set to study people's attitudes toward climate change, which of the following do you think would pose a challenge for you?

- Fine-grained estimation of the demographic attributes of all SMC users in the data set is not always possible.
- It is unclear how these people behave on other online platforms or in offline life.
- Operationalising a person's attitude towards climate change from their SMP posts is non-trivial and challenging.
- **All of the above**
- None of the above

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Nonrepresentative

Question 1 10 / 10

Which of the following most precisely describes the age bias in the SMC data set you are working with?

- **There are more young people in the data set and they post more than older people.**
- There are more young people in the data set but they post less than older people.
- There are more old people in the data set and they post more than younger people.
- There are more old people in the data set but they post less than younger people.

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Drifting

Question 1 10 / 10

Based on your observations of the hashtag activity in the SMC data set, what do you think most likely happened during the time span of the data?

- **A new IPCC report was released, which increased the general discussion on climate change.**
- A major climate movement was organized, which increased the general discussion on climate change.
- Greta Thunberg was nominated for the Nobel Peace Prize, which increased the general discussion on climate change.

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Algorithmically confounded

Question 1 10 / 10

Suppose you want to study the virality of climate movement tweets, which of the following factors is most likely to algorithmically confound your study?

- SMC's word limit feature that only allows posts of a maximum length of 280 characters
- SMC's friend recommendation algorithm which recommends a user other users to follow that may be of most interest to them
- **SMC's content recommendation algorithm which recommends a user other users' posts that they may find most interesting**
- SMC's URL shortening feature that automatically shortens any URLs in a post

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Dirty

Question 1 10 / 10

In the SMC data set, do you observe more bot activity on the climate activist side or on the climate skeptic side?

- Activist
- **Skeptic**

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Sensitive

Question 1 10 / 10

For which of the following reasons do you consider the data to be sensitive?

- The posts contain the location of the poster
- The spread of hashtags can be used to infer connections among users
- Users' attitude toward climate change can be used to infer their stance on other issues
- **All of the above**

Correct!

Submit

Points 10 / 10 My submissions 1 / 1 Deadline Friday, 17 March 2023, 19:00 To be submitted alone Show model answer

The deadline for the assignment has passed (Wednesday, 29 March 2023, 19:00).

Big data for research

Question 1 10 / 10

Having examined the data set, identify the task that the data set will be most appropriate for.

- Estimating the percentage of climate activists versus climate skeptics in the world's population
- Identifying most popular climate-related topics on SMP in the entire year of 2024
- Exploring the correlation between educatedness and climate skepticism
- **Inspecting the sentiments reflected in climate discussions on SMP.**

Correct!

Submit